Data Mining in Precision Agriculture: Management of Spatial Information

Georg Ruß[1] and Alexander Brenning[2]

[1] Otto-von-Guericke-Universität Magdeburg, Germany
[2] University of Waterloo, Canada

Abstract. *Precision Agriculture* is the application of state-of-the-art GPS technology in connection with site-specific, sensor-based treatment of the crop. It can also be described as a data-driven approach to agriculture, which is strongly connected with a number of data mining problems. One of those is also an inherently important task in agriculture: yield prediction. The question is: can a field's yield be predicted in-season using available geo-coded data sets?

In the past, a number of approaches have been proposed towards this problem. Often, a broad variety of regression models for non-spatial data have been used, like regression trees, neural networks and support vector machines. But in a cross-validation learning approach, issues with the assumption of the data records' statistical independence keep emerging. Hence, the geographical location of data records should clearly be considered while establishing a regression model and assessing its predictive performance. This paper gives a short overview of the available data, points out in detail the main issue with the classical learning approaches and presents a novel spatial cross-validation technique to overcome the problems with the classical approach towards the aforementioned yield prediction task.

Keywords: Precision Agriculture, Spatial Data Mining, Regression, Spatial Cross-Validation.

1 Introduction

Information technology has become part of our everyday lives. Information-driven management techniques have become necessary and common in industry and services. Improvements in efficiency can be made in almost any part of businesses. This is especially true for agriculture, due to the modernization and better affordability of state-of-the-art GPS technology. Agricultural companies nowadays harvests not only crops but also growing amounts of data. These data are site specific – which is essentially why the combination of GPS, agriculture and data has been termed *site-specific crop management* (SSM). A large amount of information about the soil and crop properties enabling a higher operational efficiency is often contained in these data – appropriate techniques should therefore be applied to find this information. This is a rather common problem for which the term *data mining* has been coined. Data mining techniques aim at finding those patterns in the data that are both valuable and interesting for crop management.

Yield prediction is a specific agricultural problem commonly occurring. As early as possible, a farmer would like to know how much yield he is about to expect. The ability

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, LNAI 6178, pp. 350–359, 2010.

to predict yield used to rely on farmers' long-term knowledge of particular fields, crops and climate conditions. However, this knowledge is assumed to be available in the data collected during normal farming operations throughout the season(s) [23]. A multitude of sensor data are nowadays collected, measuring a field's heterogeneity. These data are fine-scale, often highly correlated and carry spatial information which must not be neglected.

The problem of yield prediction encountered can be treated as a problem of data mining and, specifically, multi-variate regression. This article will serve as a reference of how to treat a regression problem on spatial data with a combination of classical regression techniques using a novel data sampling idea. Furthermore, this article will serve as a continuation of [19]: in the previous article, the spatial data were treated with regression models which do not take the spatial relationships into account. In the present work, we will adapt existing approaches for error estimation using spatial cross-validation approaches [4,5] to the context of crop yield prediction and spatial regression more generally. Resampling-based estimation methods (such as cross-validation and the bootstrap) for dependent data in general have been investigated recently in the context of time series data [7] and paired data [6].

1.1 Research Target

The main research target of this work is to improve and further substantiate the validity of *yield prediction* approaches using multi-variate regression modeling techniques. Previous work, mainly the regression work presented in [19,22], will be used as a baseline for this work. Some of the drawbacks of the previous approach will be clearly pointed out in this article. Nevertheless, this work aims to improve upon existing yield prediction models and, furthermore, incorporates a generic, yet novel spatial clustering idea into the process. Therefore, different types of regression techniques will be incorporated into a novel spatial cross-validation framework (compare [4,5]). A comparison of using spatial vs. non-spatial cross-validation will be presented.

1.2 Article Structure

This article will start with a brief introduction into the area of precision agriculture and a more detailed description of the available data in Section 2. This will be followed by an outline of the key techniques and the novel spatial sampling technique described in this work in Section 3. The results obtained from the modeling phase will be presented in Section 4. The article will be completed with a short conclusion in Section 5, which will also point out further lines of research.

2 Data Description

The data available in this work were collected during the growing season of 2007 on two fields south of Köthen, Germany. The data for the two fields, called *F440* and *F611*, respectively, were interpolated using kriging [24] to a grid with 10 by 10 meters grid cell sizes. F440 is geographically located around N51.68, E11.99 and has a

size of roughly 95 hectares, whereas F611 has a size of around 50 hectares and is located around N51.68, E11.85. Each grid cell represents a record with all available information. The fields grew winter wheat. Nitrogen fertilizer (N) was applied three times during the growing season. Overall, for each field there are six input attributes, accompanied by the respective current year's yield (2007) as the target attribute. In total, for the F440 field there are 6446 records, for F611 there are 4970 records.

Yield is measured in metric tons per hectare ($\frac{t}{ha}$), along the harvesting lanes (spaced 8 m apart), roughly every ten meters. The yield ranges, as well as those of the remaining attributes, are provided in Table 1. Apparent electrical soil conductivity (EC25) as a measure for a number of soil properties is acquired. Satellite or aerial image processing provides a measure of vegetation called the red edge inflection point (REIP) value, at two points into the growing season (REIP32, REIP49), according to the growing stage defined in [15]. The REIP value may also be used directly for guiding fertilizations [11]. A simplified assumption is that a higher REIP value means more vegetation. Three nitrogen fertilizer dressings are applied (N1, N2, N3, in $\frac{kg}{ha}$). In the available data, due to the fields being experimental agriculture sites, the nitrogen dressings were not temporally autocorrelated. However, this phenomenon may be considered in production sites. EC, REIP and N are measured in 10-m-intervals along the lanes which are spaced 24 meters apart.

Table 1. Statistical summary of data sets

	F440				F611			
	min	mean	median	max	min	mean	median	max
EC25	39.47	50.13	50.22	60.69	38.41	54.44	53.17	81.98
N1	50.00	63.57	70.00	70.00	42.00	65.09	68.00	70.00
N2	2.00	47.60	48.00	80.00	0.00	47.89	50.00	80.00
N3	0.00	37.98	40.00	95.00	0.00	45.61	50.00	68.00
REIP32	721.33	725.11	725.19	728.14	721.41	724.37	724.41	726.09
REIP49	724.50	727.20	727.34	729.82	721.30	727.12	727.23	729.41
YIELD07	0.49	7.37	6.89	13.92	1.32	5.42	5.51	11.88

2.1 Spatial vs. Non-spatial Data Treatment

According to [10], *spatial autocorrelation* is the correlation among values of a single variable strictly attributable to the proximity of those values in geographic space, introducing a deviation from the *independent observations* assumption of classical statistics. Given a spatial data set, spatial autocorrelation can be determined using Moran's I ([16]) or semivariograms. For the data sets used in this article, each of the attributes exhibits spatial autocorrelation. Figure 1 shows two representative experimental omnidirectional semivariograms, while the remaining attributes behave similarly. In practice, it is usually also known from the data origin whether spatial autocorrelation exists. For further information it is referred to, e.g., [8].

In previous articles using the above data, such as [20,19], the main focus was on finding a suitable regression model to predict the current year's yield sufficiently well. However, the used regression models, such as neural networks [20,21] or support vector

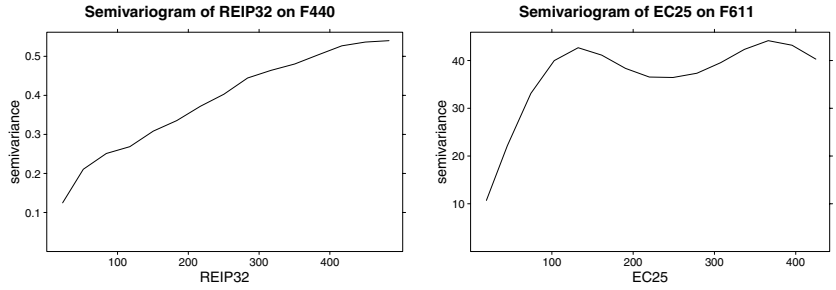


Fig. 1. Semivariograms (omnidirectional, experimental) for REIP32 (F440) and EC25 (F611)

regression [19], among others, generally assume statistical independence of the data records. However, with the given geo-tagged data records at hand, this is clearly not the case, due to (natural) spatial autocorrelation. Therefore, the spatial relationships between data records have to be taken into account, which the following section will deal with.

3 Regression Techniques on Spatial Data

Due to the shortcomings in classical regression and cross-validation learning approaches when using them on spatial data, this section will present a novel regression model for data sets which exhibit spatial autocorrelation. In non-spatial regression models, data records which appear in the training set are not supposed to appear in the test set during a cross-validation learning setup. Classical sampling methods do not take spatial neighborhoods of data records into account. Therefore, the above assumption may be rendered invalid when using non-spatial models on spatial data. This inevitably leads to overfitting and underestimates the true prediction error of the regression model (compare [4,6] for similar observations in a classification context). Therefore, the main issue is to avoid having neighboring or the same samples in training and testing data subsets during a cross-validation learning approach. The basic idea therefore is to apply changes to the resampling method and keep the regression modeling techniques as-is. The resulting procedure can be seen as spatial cross-validation technique.

3.1 From Classical to Spatial Cross-Validation

Traditionally, k-fold cross-validation for regression randomly subdivides a given data set into three parts: a training set, a validation set and a test set. A 10- to 20-fold cross-validation is usually considered appropriate to remove bias [13]. The regression model is trained on the training set until the prediction error on the validation set starts to rise. Once this happens, the training process is stopped and the error on the test set is reported for this fold. This procedure is repeated r times to remove a possible sampling bias. In our case, r has been empirically determined as 100.

In spatial data, due to spatial autocorrelation, almost identical data records may end up in training, validation and test sets. In essence, the model overfits the training data and returns an overoptimistic (biased) estimation of the prediction error. Therefore, one possible solution might be to ensure that only a very small number (if any) of neighboring and therefore similar samples end up in training and test subsets. This may be achieved by adapting the sampling procedure for spatial data. Once this issue has been accommodated, the cross-validation procedure may continue in the usual way.

3.2 Employing Spatial Clustering for Data Sampling

Given the data sets F440 and F611, a regular tessellation using a grid-based approach may be used to subdivide the fields into spatially disjunct areas. However, even though the data have been sampled on a regular grid, there are irregularities in the field. These are due to the fields' outer shape, "holes" in the data (power poles, buildings etc.) or the lanes of machinery, among other reasons. This would lead to some grid cells being much less equally dense populated than others. Therefore, a grid-based approch is rather rigid and would have to be adapted manually for each field. Hence, a more flexible method will be used here.

We assume that a spatial clustering procedure can be employed to subdivide the fields into spatially disjunct clusters or zones. The clustering algorithm can easily be run on the data records' spatial map, using the data records' longitude and latitude. Depending on the clustering algorithm parameters, this results in a tesselation map which does not consider any of the attributes, but only the spatial neighborhood between data records. A depiction of this clustering process can be found in Figure 2(a). As may be seen in the figure, the clustering leads to clusters (spatial areas) covering roughly the same number of points, due to the relatively regular data point density encountered here. In analogy to the non-spatial regression treatment of these data records, now a spatially-aware cross-validation regression problem can be handled using the k resulting zones of the clustering algorithm as an input for k-fold cross-validation. This ensures that the training set has only a small amount of spatial autocorrelation with the test set. Standard models, as described below, can be used straightforwardly, without requiring changes to the models themselves. The experimental setup and the results are presented in the following section.

The training and test sets are selected from the clusters using random sampling. Therefore, a small number of points in neighboring areas are still possibly spatially autocorrelated. This could be avoided by using a sampling method which takes the spatial relationships between the clusters into account. However, when comparing the standard, non-spatial regression setup to the one described here, it is assumed that the difference in the error underestimation is much higher than the one of introducing a space-aware sampling method on the clusters.

The spatial clustering procedure may be considered as a broader definition of the standard cross-validation setup. This can be seen as follows: when refining the clustering further, the spatial zones on the field become smaller. The border case is reached when the field is subdivided into as many clusters as there are data records, i.e. each data record describes its own cluster. In this special case, the advantages of spatial clustering are lost since no spatial neighborhoods are taken into account in this approach.

Therefore, the number of clusters should be seen as a tradeoff between predictive precision and statistical validity of the model. The parameter k for the size of the tessellation has to be determined heuristically.

3.3 Regression Techniques

In previous work ([19,20]), numerous regression modeling techniques have been compared on similar data sets to determine which of those modeling techniques works best. Support vector regression has been determined as the best modeling technique. It has furthermore recently been shown to work rather successfully in spatial classification tasks, albeit without spatial cross-validation, as in [17]. Hence, in this work support vector regression will serve as a benchmark technique against which further models will have to compete. Experiments are conducted in R [18]. It is assumed that the reader is mostly familiar with the regression techniques below. Therefore, the techniques used are described in short. References to further details are given, where appropriate. The performance of the models will be determined using the root mean squared error (RMSE).

Support Vector Regression. Support Vector Machines (SVMs) are a supervised learning method discovered by [1]. They were originally described for the use in classification, but can also be applied to regression tasks, where optimization of a cost function is achieved. The model produced by support vector regression depends only on a subset of the training data – which are essentially the support vectors. Further details can be found in [19]. In the current experiments, the *svm* implementation from the *e1071* R package has been used.

Regression Trees. Regression trees have seen some usage in agriculture [9,12,14]. Essentially, they are a special case of decision trees where the outcome (in the tree leaves) is a continuous function instead of a discrete classification. The *rpart* R package has been used.

Random Forests. According to [3], random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. In the version used here, the random forest is used as a regression technique. Basically, a random forest is an ensemble method that consists of many regression trees and outputs a combined result of those trees as a prediction for the target variable. Usually, the generalization error for forests converges to a limit as the number of trees in the forest becomes large. The *randomForest* R package has been used.

Bootstrap Aggregating. Bootstrap aggregating (or bagging) has first been described in [2]. It is generally described as a method for generating multiple versions of a predictor and using these for obtaining an aggregate predictor. In the regression case, the prediction outcomes are averaged. Multiple versions of the predictor are constructed by taking bootstrap samples of the learning set and using these as new learning sets. Bagging is generally considered useful in regression setups where small changes in the training data set can cause large perturbations in the predicted target variables. Since random forests are a variant of bagging where regression trees are used as the internal predictor, both random forests and bagging should deliver similar results. Running them on the data sets should therefore deliver similar

results as well, since the bagging implementation in the R *ipred* package internally uses regression trees for prediction. Therefore, the main difference between random forests and bagging in this article is that both techniques are implicitly run with different parameters.

4 Results

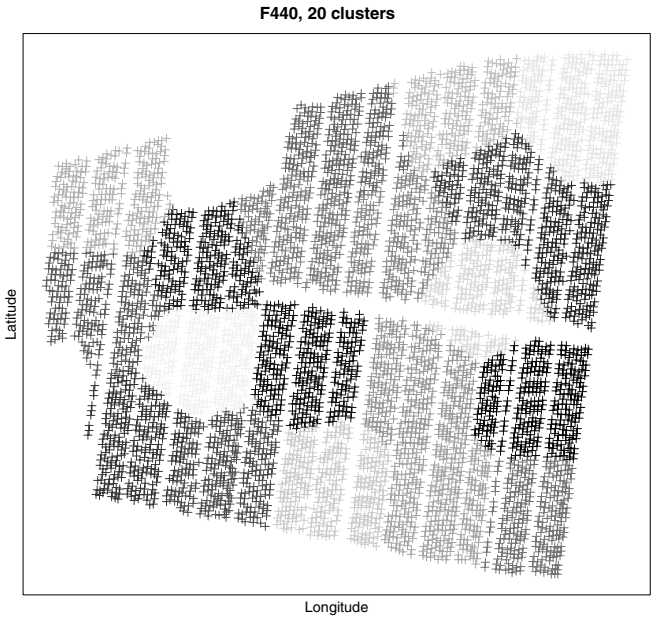
The main research target of this article is to assess whether existing spatial autocorrelation in the data sets may fail to be captured in standard, non-spatial regression modeling setups. Therefore, the approach consists of a comparison between a non-spatial and a spatial regression with cross-validation setup. The non-spatial setup was described in Section 3.1, the spatial setup has been presented in Section 3.2.

The results in Table 2 confirm that the spatial autocorrelation inherent in the data set leads classical, non-spatial regression modeling setups to a substantial underestimation of the prediction error. This outcome is consistent throughout the results, regardless of the used technique and regardless of the parameters. Furthermore, it can be seen that Random Forests seem to yield better performance in terms of lower prediction error, regardless of the setup used. For an illustrative depiction of the RMSE in the spatial approach see Figure 2(b), which shows the dataset partitioned into 50 spatial clusters with the cross-validation RMSE displayed.

Moreover, the spatial setup can be easily set to emulate the non-spatial setup: set k to be the number of data records in the data set. Therefore the larger the parameter k is set, the smaller the difference between the spatial and the non-spatial setup should be. This assumption also holds true for almost all of the obtained results.

Table 2. Results of running different setups on the data sets F440 and F611; comparison of spatial vs. non-spatial treatment of data sets; root mean squared error is shown, averaged over clusters/folds; k is either the number of clusters in the spatial setup or the number of folds in the non-spatial setup

		F440		F611	
	k	spatial	non-spatial	spatial	non-spatial
Support Vector Regression	10	1.06	0.54	0.73	0.40
	20	1.00	0.54	0.71	0.40
	50	0.91	0.53	0.67	0.38
Regression Tree	10	1.09	0.56	0.69	0.40
	20	0.99	0.56	0.68	0.42
	50	0.91	0.55	0.66	0.40
Random Forest	10	0.99	0.50	0.65	0.41
	20	0.92	0.50	0.64	0.41
	50	0.85	0.48	0.63	0.39
Bagging	10	1.09	0.59	0.66	0.42
	20	1.01	0.59	0.66	0.42
	50	0.94	0.58	0.65	0.41



(a) k-means clustering on F440, $k = 20$

(b) spatial cross-validation on field F440, $k = 50$, RMSE is shown

Fig. 2. k-means clustering on F440 and resulting cross-validation RMSE

5 Conclusions and Future Work

This article elaborated upon a central data mining task: regression. Based on two data sets from precision agriculture, a continuation and improvement over previous work ([19,20]) could be achieved. The important difference between spatial data and non-spatial data was pointed out. The implications of spatial autocorrelation in these data sets were mentioned. From an information management point of view, neighboring data records in a spatially autocorrelated data sets are not supposed to end up in training and test sets since this leads to a considerable underestimation of the prediction error, regardless of the used regression model.

It can be concluded that it is indeed important to closely consider spatial relationships inherent in the data sets. As a suggestion, the following steps should be taken: for those data, the spatial autocorrelation should be determined. If spatial autocorrelation exists, standard regression models must be adapted to the spatial case. A straightforward and illustrative approach using simple k-means clustering has been described in this article.

5.1 Future Work and Acknowledgements

Despite having improved and validated upon the yield prediction task, the data sets carry further information. *Variable importance* refers to the question which of the variables is actually contributing most to the yield prediction task. *Management zones* refers to discovering interesting zones on the (heterogeneous) field which should be managed differently from each other.

The data in this work have been obtained on the experimental farm Görzig and were acquired from Martin Schneider and Peter Wagner from Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für landwirtschaftliche Betriebslehre.

References

1. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pp. 144–152. ACM Press, New York (1992)
2. Breiman, L.: Bagging predictors. Technical report, Department of Statistics, Univ. of California, Berkeley (1994)
3. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
4. Brenning, A.: Spatial prediction models for landslide hazards: review, comparison and evaluation. Natural Hazards and Earth System Science 5(6), 853–862 (2005)
5. Brenning, A., Itzerott, S.: Comparing classifiers for crop identification based on multitemporal landsat tm/etm data. In: Proceedings of the 2nd workshop of the EARSeL Special Interest Group Remote Sensing of Land Use and Land Cover, pp. 64–71 (September 2006)
6. Brenning, A., Lausen, B.: Estimating error rates in the classification of paired organs. Statistics in Medicine 27(22), 4515–4531 (2008)
7. Bühlmann, P.: Bootstraps for time series. Statistical Science 17, 52–72 (2002)
8. Cressie, N.A.C.: Statistics for Spatial Data. Wiley, New York (1993)
9. Crone, S.F., Lessmann, S., Pietsch, S.: Forecasting with computational intelligence - an evaluation of support vector regression and artificial neural networks for time series prediction. In: International Joint Conference on Neural Networks, 2006. IJCNN '06, pp. 3159–3166 (2006)

10. Griffith, D.A.: Spatial Autocorrelation and Spatial Filtering. In: Advances in Spatial Science, Springer, New York (2003)
11. Heege, H., Reusch, S., Thiessen, E.: Prospects and results for optical systems for site-specific on-the-go control of nitrogen-top-dressing in germany. Precision Agriculture 9(3), 115–131 (2008)
12. Huang, C., Yang, L., Wylie, B., Homer, C.: A strategy for estimating tree canopy density using landsat 7 etm+ and high resolution images over large areas. In: Proceedings of the Third International Conference on Geospatial Information in Agriculture and Forestry (2001)
13. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of International Joint Conference on Artificial Intelligence (1995)
14. Lobell, D.B., Ortiz-Monasterio, J.I., Asner, G.P., Naylor, R.L., Falcon, W.P.: Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. Agronomy Journal 97, 241–249 (2005)
15. Meier, U.: Entwicklungsstadien mono- und dikotyler Pflanzen. In: Biologische Bundesanstalt fünd- und Forstwirtschaft, Braunschweig, Germany (2001)
16. Moran, P.A.P.: Notes on continuous stochastic phenomena. Biometrika 37, 17–33 (1950)
17. Pozdnoukhov, A., Foresti, L., Kanevski, M.: Data-driven topo-climatic mapping with machine learning methods. Natural Hazards 50(3), 497–518 (2009)
18. R Development Core Team: R: A Language and Environment for Statistical Computing. In: R Foundation for Statistical Computing, Vienna, Austria (2009) ISBN 3-900051-07-0
19. Ruß, G.: Data mining of agricultural yield data: A comparison of regression models. In: Perner, P. (ed.) Advances in Data Mining. Applications and Theoretical Aspects. LNCS, vol. 5633, pp. 24–37. Springer, Heidelberg (2009)
20. Ruß, G., Kruse, R., Schneider, M., Wagner, P.: Estimation of neural network parameters for wheat yield prediction. In: Bramer, M. (ed.) AI in Theory and Practice II, July 2008. Proceedings of IFIP 2008, vol. 276, pp. 109–118. Springer, Heidelberg (July 2008)
21. Ruß, G., Kruse, R., Schneider, M., Wagner, P.: Optimizing wheat yield prediction using different topologies of neural networks. In: Verdegay, J., Ojeda-Aciego, M., Magdalena, L. (eds.) Proceedings of IPMU '08, pp. 576–582. University of Málaga (June 2008)
22. Ruß, G., Kruse, R., Wagner, P., Schneider, M.: Data mining with neural networks for wheat yield prediction. In: Perner, P. (ed.) ICDM 2008. LNCS (LNAI), vol. 5077, pp. 47–56. Springer, Heidelberg (2008)
23. Stafford, J.V., Ambler, B., Lark, R.M., Catt, J.: Mapping and interpreting the yield variation in cereal crops. Computers and Electronics in Agriculture 14(2-3), 101–119 (1996), Spatially Variable Field Operations
24. Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging. Springer Series in Statistics. Springer, Heidelberg (June 1999)